APPROACHES TO ELECTRONIC PROCESSING OF ARCHIVAL MASS SOURCES: REFORMATTING OR CREATING DIGITAL RESOURCES1

Lyudmila Mazur and Oleg Gorbachev

Abstract: There are two approaches commonly applied in research to digitize primary data sources: problem-orientated and source-orientated. The first one is more economical, efficient and flexible. What we get, however, is a single-use resource. The second approach is aimed at designing a system that would contain the complete set of authentic data. Thus, the system would be suitable for different purposes, but it would take much more time to construct.

This paper analyzes a case of two sets of primary data and the problems faced when designing and creating a digital multi-purpose resource system. The first data set includes primary forms of the All-Russia Communist Party Census of 1922 and 1927. The second set comprises family budget records taken during budget surveys of 1929 and 1963. Both sources have a complex hierarchical structure, difficult to adjust for data normalization. Finally, the records taken in different years characterize the dynamics for specific social groups, which presents us with the challenging task of organizing this information and solving the problem of incompatibility between the structure of the records. All the above-mentioned characteristics can lead to the loss of primary data when converting them into a database.

When creating a digital resource system, it is essential to develop the right digitization strategy, one that is able to compensate for data loss and able to support a system suitable for a variety of purposes and adjustable to the individual needs of each user.

Keywords: Reformatting, digital resources, digital archive, database, budget surveys, Communist Party Census

Today there is a common understanding (more or less) of how to digitize ancient manuscripts: a complete copy is usually required, with all the

[1] The research is supported by Russian Foundation of Basic Research, project No. 18-09-00592 "The Evolution of the Peasant Family in the Middle Urals in the 20th Century: Reconstruction Based on Budgetary Surveys."

original features of the document, but there is no universally accepted approach to documents of the 20th century. They have not yet been covered by the noble patina of the past. And very often at first sight they contain an excessive amount of information. What if the volume of the information in the document significantly exceeds the immediate needs of the researcher?

The formation of the digital information environment is one of the trends in the development of the discipline of history, which, along with other humanities, is undergoing a digital change and faces a lot of longstanding unresolved problems. We'd like to begin with outlining the main trends of the development of history as a discipline:

– *thematically*, from the middle of the 20th century, there has been a transition from the history of events to the study of processes and history of everyday life, mentality, memory, etc.

– *methodologically,* the zone of interdisciplinarity is expanding, which affects not only the conceptual apparatus but also the formulation of research issues and the use of methods from related sciences, especially sociology, economics, statistics, geography, and so on.

– *from the point of view of the information and resource potential*, there is an expansion of the base of available sources for historical research to include mass sources, primarily ego-documents, nominative sources, periodicals and so on.

– *technologically,* there is a total computerization of historical research practices, because today the computer is indispensable in a historian's work, although the scope of its use can vary significantly: from searching for information and composing text to using the capabilities of computer technologies at all stages of the research process, including the processing and transformation of historical data with the help of a variety of standard or specialized program packages – databases, GIS, media and so on. As a significant part of the professional community of historians begins to actively use digital resources, there comes a wide recognition of the fact that digital orientation in the expanding world of electronic resources becomes part of the "historian's craft", according to Leonid Borodkin (Borodkin).

An integral element of all these directions is *the creation of digital resources* (in Russian, *digitalizatsia*). It should be distinguished from *digital reformatting*, which means transferring information from physical to digital

media, such as scanning a document. Reformatting is now realized in the archival sector. In Russia, according to the Program for the Informatization of Archives, by 2020, 20% of all documentary arrays should be reformatted. In the process of reformatting, there is no change in the structure of information: it simply acquires an electronic form, convenient for use in digital format. This approach can be correlated with the third industrial revolution, which lasted from 1969 to 2010.

But the creation of an electronic analogue of a paper document does not fundamentally change the information environment and only facilitates remote access to archival documents. This procedure is appropriate when working with unique texts. If we are dealing with complexes of mass nominative sources, then an intermediate step in the process of information transfer is its formalization and transformation into data structures available for additional quantitative or spatial processing.

The creation of digital resources is understood more widely – this is the process of creating a new information product in digital form. For example, the creation of databases on census materials, electronic maps, 3D-reconstructions and so on. In this case, the information of the primary source is transformed into a new form with the inevitable change in the initial information potential (its expansion or contraction). The key advantage of such work is the creation of a new innovative product, with new functional and consumer properties. And if *reformatting* is primarily aimed at improving existing traditional research practices, *the creation of digital resources* allows us to bring them to a new level of interdisciplinary interaction and use. On this basis, the creation of digital resources can be considered a phenomenon of the fourth industrial revolution.

The Russian specificity of the process reveals two main trends:

- in the archival sphere, most digital transformation initiatives are aimed at *reformatting*;

- in the discipline of history, primarily in digital history, the creation of new information products prevail. However, the possibilities of using them are limited, since they presuppose the creation of information products intended for a specific task. As a rule, the secondary use of these products is difficult or impossible.

The most important task of developing the infrastructure of historical research is the creation of open information resources, which are the result

of *creating digital resources*. Requirements for the development of resources of this kind include the following criteria:

1. *Source-orientation*. The creation of a digital resource assumes the measurement, formalization, normalization of primary materials of mass historical sources (individual forms of population censuses, metric records, *revizskiye skazki*, and other variants of population registers), which contain mainly personal (nominative) information.

2. *Completeness*. The resource includes the entire array of surviving documents or records.

3. *Efficiency*, i.e. resource uses a data format that provides storage, import, integration of non-recurring resources and their collective use.

4. *Openness*. The resource is available to the scientific community.

The ideology of creating source-oriented resources available online was developed in the concept of the "big database". It is built on the observance of certain standards for the description of data and the use of such storage formats that ensure their viability under conditions of rapid change of hardware and software (Gorbachev).

The implementation of international projects for the creation of big databases (IPUMS-USA, NAPP, Mosaic, the Vienna Database on European Family History, church books, the Demographic Data Base at Umeå University, etc.) demonstrates the high scientific potential of this scientific trend, which contributes to the formation of a new research paradigm that supplants the national frameworks for research practices and creates opportunities for cross-national research.

There are two main strategies for creating digital resources of primary mass sources: a *research-oriented* and *source-oriented approach*. The former implements the principle of selecting information from a source in accordance with the question being studied. It is most often used because it is time-saving, operative and adaptive, considers the research requests and hypotheses. But in the end, a one-time resource is created.

The latter is aimed at maximally complete and authentic representation of source information in the system being designed. In this sense, it corresponds to the tasks of multi-purpose use, but it is more labor-intensive, especially in the case of information-intensive and complexly organized sources. These are, for example, the primary forms of the All-Russian Communist Party Census of 1922, which include 59 questions, many of which contain

sub-questions. After normalization, the number of questions in the form has grown to 120.

Another example of a complex structured source is the primary forms of budgetary surveys of peasant households, whose information potential reaches several thousand features (about 5,000). Implementing a source-oriented approach to them is almost impossible: loss of information is inevitable. What to do in this case?

Thus, one of the crucial tasks of creating a multipurpose information resource is the development of a strategy for creating digital resources that would allow to ensure a mode of multipurpose use of the resource with compensation of information losses and flexible adaptation to individual requests of different users.

As an approach to creating digital resources, we offer a combination of a database and an electronic archive of document scans. This strategy was tested using the example of the two sets of sources mentioned above: the first complex includes the primary forms of the All-Russian CP Census of 1922 and 1927, and the second complex are the household budgets, preserved as a result of budgetary surveys in 1928/29 and 1963.

Both sources are characterized, first, by a complex hierarchical structure, poorly adapted to normalize the data; secondly, they contain a large amount of primary data describing the object (party member or peasant household). Thirdly, they represent a dynamic complex, including descriptions of the object from different years. All these features cause loss of primary information when converted to a database.

Let us dwell on their characteristics.

I. When designing the Information system "All-Russian CP census 1922–1923", the main question was how to implement the project: to follow the structure and features of a source or to create an information resource according to the concrete research task? (Taller) The second option is in a sense simpler and more efficient since the problem of formalizing the source information is also solved, taking into account the requirements and tasks formulated by the researcher. However, the choice was made in favor of *a source-oriented model* that allows the reuse of data for different research purposes by maximizing the full display of information contained in a complex of sources (Gutnov & Pereverten'). Source orientation manifests itself in the structure of the information array, which includes a database with

structured but not formalized questionnaire information and electronic copies of documents (questionnaires) in JPEG format. Electronic copies are tied to specific records – i.e. the user, working with the archive, can study the information entered into the database and simultaneously view the electronic copy of the questionnaire by checking or specifying information.

The need to include electronic copies of documents in the archive is justified by the following considerations:

1. *The needs of control over the correctness of data entry*. This control can additionally be carried out by the user, comparing information in the database with a copy of the document;

2. *The possibilities of reorganizing the database*. Despite the declared principle of completeness in the development of the structure of the database, it was impossible to structure and reflect all the information available in the questionnaires. Something is inevitably lost or does not fit into the database format, but can be interesting to the user. In this case, it is necessary to create conditions for the user to "refine" the database for their tasks;

3. *Visualization of the source allows us to clarify its features*: the appearance, the nature of the records, additional remarks on the questionnaire, the correctness of its design – this is especially important in cases of illegible handwriting of the registrar.

The type of this electronic resource can be defined as *prosopographical* since it has the characteristic features caused by a complex of sources. For example, the database contains personal data of communists not only at the time of the census but also in dynamics concerning labor and party activities and military service. Generalization of this information makes it possible to compile a collective biography of the members of the Ural organizations of the Russian Communist Party (Bolsheviks) of the 1920s.[2]

The information system operates in two modes: search and browsing. In the first case, the user can search for a name or the name of the locality to find information about a person. The second mode is working with filters to sample records according to a given criterion (age, nationality, education, social status, etc.) to use statistical methods to obtain indicators for both creating a collective social portrait of the communists of the Urals from

[2] See the interpretation of prosopography in the framework of historical informatics in: Yumasheva.

1922 to 1924 and for the study of individual groups of communists. Filters can be used not only in database mode; tables can be imported into a more flexible and user-friendly format (for example, Excel) and be processed.

The database "All-Russian Census of Members of the RCP(b): Ekaterinburg Province, 1922" consists of seven tables in accordance with the main thematic blocks of Form A: general information (14 fields), education (18 fields), social origin (9 fields), labor activity (12 fields), party activities (15 fields), revolutionary activity (17 fields) and military service (27 fields). The tables are linked through the key fields "Sequence number" and "Last name, first name, middle name." The size of the database is 12,000 records, plus about 10,000 for the set from 1924. The database is posted on the website of the project "Early Soviet Society As a Social Project: Ideas, Implementation Mechanisms, Design Results", currently test mode. A copy of the database was provided to the Public Organizations Documentation Center of the Sverdlovsk Region for use as a search tool.

II. Another resource, the creation of which began in 2018, includes information on primary forms of budgetary surveys of peasant farms from 1928 to 29 and in 1963.

Methods for organizing and carrying out budgetary surveys were developed in the Russian Empire at the end of the 19th century, thanks to the activity of zemstvo (a type of regional representative authority). After the Bolshevik Revolution of 1917, these investigations were renewed and originally acquired the form of a periodic analysis, carried out in cluster surveys, covering mainly peasants' households.

In 1921 and 1922, about 500 households were surveyed; in 1922 and 1923, 4,000. Between 1923 and 1930, the Central Statistics Office surveyed from 8,000 to 20,000 peasant households annually (Istoriya obsledovaniy…). The scope of these surveys reached its peak just before the beginning of the collectivization campaign. In 1928 and 1929, in the USSR, budget surveys covered about 20,000 households (Bokarev).

Peasant households were selected using the cluster method: at first, specialists specified the number of households for each region; then this region was divided into production areas (specializing in crop growing, cattle farming, timber production, fishery, pasture farming and so on). This approach gave statisticians the full picture of the life of peasant households and enabled them to study the impact of macroeconomic factors on family budgets.

Beginning in 1932, a permanent network of households was formed that included around 0.01% of the USSR's population (about 6,500 families). From 1969 to 1987, budget surveys covered 62,000 households (Istoria obsledovaniy…). Since the 1970s, in many parts of the country, including Sverdlovsk region, the surveys started to focus on families of *sovkhoz (state farms)* workers, which replaced *kolkhoz* families.

The data from the budgetary survey was aggregated and was found to contain the following sections: common household data with sizes and composition; working hours; income obtained from work (in a collective farm, at an enterprise or in an institution); turnover of products in a household; expenses on acquisition of industrial goods, transport, housing, household services, taxes and debts; and food structure.

The most complete data is the *primary* forms of budgetary surveys, which were completed for each farm.

In the State Archives of the Sverdlovsk Region, 325 budgets were found for 1928–1929, including several thousand data points, as well as 221 "Control Writing Books of the Statistician" for 1963, which allows for a comparative analysis. Meanwhile, the budgetary inspections allow us to reconstruct the life of a household not only demographically but also in terms of its budget and consumption.

Based on the comparable indicators of 1928–29 and 1963, a dynamic database is being formed to study the evolution of the peasant family in the Urals from the 1920s to 1960s.

The database being created implements a strategy that is oriented towards creating a *research*-oriented database but with the possibility of modifying it for secondary use. The basic structure of the table includes 43 fields, most fully revealing the size, composition, type of family and factors influencing its demographic characteristics. They can be divided into the following blocks:

Demographic, such as the household head's profile (age, sex, nationality, literacy, affiliation with public organizations), data on the family size and structure, the age of household members, the number of minor children and the demographic family type

Economic, such as the year when the household was created; the land, cattle and other livestock; farming practices such as growing grain and vegetable crops and haymaking; crafts; transport and living conditions

Budget-related, such as income and expenditures, consumption patterns and valuable possessions

The database reflects approximately 4% of the information stored in the primary sources. Therefore, in parallel with the database, an e-archive of electronic copies of archival documents is formed. Each record of the database is associated with the corresponding file in the e-archive. Creating an open resource, we are striving to ensure that the user, relying on the e-archive, could make corrections or additions to the database while performing research.

Returning to the requirements for creating an open information resource based on historical mass sources, especially nominative sources, it should be noted that the requirements are fully feasible only when working with sources that include a relatively small set of data. If the source has a more complex structure and contains a large amount of primary information, it is possible to create a database including part of the primary source, with the ability for users to correct and enhance the data.

References

Bokarev, Yu. P. *Byudzhetnye obsledovaniya krest'yanskikh khozyaystv 20-kh godov kak istoricheskiy istochnik (Budget surveys of peasant farms of the 1920s as a historical source).* Moscow : Nauka, 1981.

Borodkin, L. I. "Digitalizatsiya, vizualizatsiya, reprezentatsiya, analitika?" (Digitalization, visualization, representation, analytics?), *Informatsionnyy byulleten' assotsiatsii 'Istoriya i komp'yuter'* 44 (2015), 3–8.

Digitalizatsia (Digitalization) // Entsiklopedicheskiy slovar' SMI [Electronic resource]. URL: https://smi.academic. https://smi.academic.ru/ 74/%D0% 94%D0%B8%D0% B3%D0%B8%D1%82%D0%B0%D0%BB%D0%B8 %D0 %B7%D0%B0%D1%86% D0%B8% D1%8F (addressed 03.03.2019).

Istoriya obsledovaniy byudzhetov domashnikh khozyaystv (History of household budget surveys) [Electronic resource]. URL: http://stavstat.gks.ru/wps/wcm/connect/rosstat_ts/stavstat/resources/f233f100415db7128daaef-367ccd0f13/История+обследований+бюджетов+домашних+хозяйств.pdf. (addressed 03.03.2019).

Gorbachev, O. V. "*Istoriko-demograficheskie bazy dannykh v kontekste evropeyskikh komparativnykh issledovaniy: proekt EHPS-NET" (Historical and demographic databases in the context of European comparative research: the EHPS-NET project),* Istoricheskaya informatika. Informatsionnye tekhno-

logii i matematicheskie metody v istoricheskikh issledovaniyakh i obrazovanii. 2014. No. 2-3 (8-9). Pp. 4–9.

Gutnov, D. A. and Pereverten', V. A. "Rossiyskie istoriki XVIII – nachala XX vv.: proekt i informatsionnaya Sistema" (Russian historians of the 18th - early 20th centuries: project and information system). in *Krug idey: novoe v istoricheskoy informatike*. Moscow, 1994, 49–50.

Thaller, M. "Chto takoe 'istochniko-orientirovannaya obrabotka dannykh'; chto takoe 'istoricheskaya informatika'" (What is the "source-oriented data processing"; what is "historical informatics"). in *Istoriya i komp'yuter: novye informatsionnye tekhnologii v istoricheskikh issledovaniyakh i obrazovanii*. Goettingen, 1993, 5–18.

Yumasheva, Yu. Yu. "Istoriografiya prosopografii" (Historiography of prosopography). *Izvestiya Ural'skogo gosudarstvennogo universiteta.* Seriya 2. Gumanitarnye nauki. 10/39/ (2005), 95–127.

About the authors:

Mazur Lyudmila N., Doctor of Historical Sciences and Professor at the Ural Federal University in Ekaterinburg (Russia), the Director of International Demographic Unit. *E-mail:* Lmaz@mail.ru

Gorbachev Oleg V., Doctor of Historical Sciences and Professor at the Ural Federal University in Ekaterinburg (Russia). E-mail: og_06@mail.ru

Резюме

Существует два подхода, обычно применяемых в исследовательской практике для оцифровки первичных источников данных: проблемно-ориентированный и источнико-ориентированный. Первый более экономичный, эффективный и гибкий. Однако в этом случае мы получаем одноразовый ресурс. Второй подход направлен на разработку системы, содержащей полный набор достоверных данных. Таким образом, систему можно будет использовать для различных исследовательских целей, но для ее создания потребуется гораздо больше времени.

В этой статье анализируются два комплекса первичных данных и проблемы, с которыми мы столкнулись в ходе проектирования и создания многоцелевого цифрового ресурса. Первый комплекс включает в себя первичные формы Всероссийских партийных переписей 1922 и 1927 гг. Второй содержит данные семейных бюджетов, собранные в ходе бюджетных обследований 1929 и 1963 гг. Оба источника имеют сложную иерархическую структуру, затрудняющую нормализацию. Записи, сделанные в

разные годы, позволяют проследить динамику для конкретных социальных групп, но для этого необходимо решить сложную задачу организации информации с целью достижения совместимости структур первичных данных. Существует опасность потери первичной информации при преобразовании в единую базу данных.

При создании системы цифровых ресурсов важно разработать правильную стратегию оцифровки, которая может компенсировать потерю данных и создать систему, которая подходит для различных целей и может быть адаптирована к индивидуальным потребностям каждого пользователя.